11
Evaluation and Technology Assessment

CHARLES P. FRIEDMAN, JEREMY C. WYATT, AND DOUGLAS K. OWENS [JW1]

After reading this chapter, you should know the answers to these questions:

- Why are empirical studies based on the methods of evaluation and technology assessment important to the successful implementation of information resources to improve health care?
- What challenges make studies in informatics difficult to carry out? How are these challenges addressed in practice?
- Why can all evaluations be classified as empirical studies?
- What are the major assumptions underlying objectivist and subjectivist approaches to evaluation? What are the advantages and disadvantages of each?
- What are the factors that distinguish the three stages of technology assessment?
- How does one distinguish measurement and demonstration aspects of objectivist studies, and why are both aspects necessary?
- What steps are typically undertaken in a measurement study? What designs are typically used in demonstration studies?
- What is the difference between cost-effectiveness and cost-benefit analyses? How can investigators address issues of cost effectiveness and cost benefit of medical information resources?
- What steps are followed in a subjectivist study? What techniques are employed by subjectivist investigators to ensure rigor and credibility of their findings?
- Why is communication between investigators and clients central to the success of any evaluation?

11.1 Introduction and Definitions of Terms

This chapter is about the formal study of medical information resources—computer systems that support health care, education, research, and biomedical research—to address questions of importance to developers, users, and other people. We explore the methods of performing such studies, which are essential to the field of informatics but are often challenging to carry out successfully. Fortunately, every study is not designed from a blank tablet. To guide us, there exist two closely related and highly overlapping bodies of methodological knowledge: evaluation and technology assessment. These methodological fields, which have largely developed over the past four decades, are together the subject of this chapter.[1]

[1]This chapter is heavily drawn from the textbook on evaluation by co-authors Friedman and Wyatt (1997a); refer to that text for further details.

11.1.1 Evaluation and Technology Assessment

Most people understand the term evaluation to mean a measurement or description of an organized, purposeful activity. Evaluations are usually conducted to answer questions or to help make decisions. Whether we are choosing a holiday destination or a word processor, we evaluate what the options are and how well they fit key objectives or personal preferences. The forms of the evaluation differ widely, according to what is being evaluated and how important the decision is. Thus, in the case of holiday destinations, we may ask our friend which Hawaiian island she prefers and may browse color brochures from the travel agent; for a word processor, we may gather technical details, such as the time to open and spell check a 1,000-word document or the compatibility with our printer. Thus, the term **evaluation** describes a wide range of data-collection activities, designed to answer questions ranging from the casual, "What does my friend think of Maui?" to the more focused, "Is word processor A faster than word processor B on my personal computer?"

In medical informatics, we study the collection, processing, and communication of health care information and build **information resources**—usually consisting of computer hardware or software—to facilitate these activities. Such information resources include systems to collect, store, and retrieve data about specific patients (e.g., clinical workstations and databases) and systems to assemble, store, and reason about medical knowledge (e.g., medical knowledge–acquisition tools, knowledge bases, decision-support systems, and intelligent tutoring systems). Thus, there is a wide range of medical information resources to evaluate.

Further complicating the picture, each information resource has many different aspects that can be evaluated. The technically minded might focus on inherent characteristics, asking such questions as, "Is the code compliant with current software engineering standards and practices?" or "Is the data structure the optimal choice for this type of application?" Clinicians, however, might ask more pragmatic questions such as, "Is the knowledge in this system completely up-to-date?" or "How long must we wait until the decision-support system produces its recommendations?" People who have a broader perspective might wish to understand the influence of these resources on users or patients, asking questions such as, "How well does this database support a clinical audit?" or "What effects will this decision-support system have on clinical practice, working relationships, and responsibilities?" Thus, evaluation methods in medical informatics must address a wide range of issues, from technical characteristics of specific systems to systems' effects on people and organizations.

Technology assessment is a field of study closely aligned with evaluation (Garber and Owens, 1994). The Institute of Medicine (1985, p. 2) defines technology assessment as "any process of examining and reporting properties of a medical technology used in health care, such as safety, efficacy, feasibility, and indication for use, cost, and cost effectiveness, as well as social, economic, and ethical consequences, whether intended or unintended."

But what is a medical technology? **Medical technology** usually is defined broadly and consists of the "techniques, drugs, equipments, and procedures used by health care professionals in delivering medical care to individuals, and the systems within which such

care is delivered" (Institute of Medicine, 1985, pp. 1–2). Medical information resources clearly fit within this definition. Technology assessment is relevant to informatics because many of the techniques from this field are applicable to the study of information resources.

We shall not dwell here on the differences between evaluation and technology assessment. Such differences are ones of emphasis and focus. Individuals who do evaluation and technology assessment are interested in much the same issues and use similar methods.

11.1.2 Reasons for Performing Studies

Like all complex and time-consuming activities, evaluation and technology assessment can serve multiple purposes. There are five major reasons why we study clinical information resources (Wyatt and Spiegelhalter, 1990):

- *Promotional*: If we are to encourage the use of information resources in medicine, we must be able to reassure physicians that these systems are safe and that they benefit both patients and institutions through improved cost effectiveness.
- *Scholarly*: One of the main activities in medical informatics is developing clinical information resources using computer-based tools. To obtain a deeper understanding of the links between the structure, function, and effects of these information resources on clinical decisions and actions requires careful evaluation. The knowledge we gain from such studies will help to build the foundations of medical informatics as a discipline (Heathfield and Wyatt, 1995).
- *Pragmatic*: Without evaluating their systems, developers will never know which techniques or methods are more effective or why certain approaches failed. Equally, other developers will not be able to learn from previous mistakes and may reinvent a square wheel.
- *Ethical*: Clinical professionals are under an obligation to practice within an ethical framework. For example, before using an information resource, health care providers must ensure that it is safe. Equally, those responsible for commissioning the purchase of a hospital-wide clinical information system costing several million dollars must be able to justify this in preference to other information resources or the many other health care innovations that compete for the same budget.
- *Medicolegal*: To reduce the risk of liability, developers of an information resource should obtain accurate information to allow them to assure users that the resource is safe and effective. Users need evaluation results to enable them to exercise their professional judgment before using systems so that the law will regard these users as "learned intermediaries." An information resource that treats users merely as automata, without allowing them to exercise their skills and judgment, risks being judged by the strict laws of product liability instead of by the more lenient principles applied to provision of professional services (Brahams and Wyatt, 1989) (also see Chapter 10).

The motivation for every study is one or more of these factors. Awareness of the major reason for conducting an evaluation will often help the investigators to frame the questions to be addressed and to avoid disappointment.

11.1.3 The Stakeholders in Evaluation Studies and Their Roles

Figure 11.1 shows the actors who pay for (solid arrows) and regulate (shaded arrows) the health care process. Each of them may be affected by a medical information resource, and each may have a unique view of what constitutes benefit. More specifically, in a typical clinical information resource project, the key stakeholders are the developers, the users, the patients whose management may be affected, and the people responsible for purchasing and maintaining the system. Each may have different questions to be answered (Figure 11.2).

Whenever we design evaluation or technology assessment studies, it is important to consider the perspectives of all stakeholders in the information resource. Because studies are often designed to answer specific questions, any one study is unlikely to satisfy all of the questions that concern stakeholders. Sometimes, due to the intricacy of health care systems and processes, it can be a challenge for an evaluator to identify all the relevant stakeholders and to distinguish those whose questions must be satisfied from those whose satisfaction is optional.

11.2 The Challenges of Study Design and Conduct

The work of evaluation and technology assessment in informatics lies at the intersection of three areas, each notorious for its complexity: (1) medicine and health care delivery,

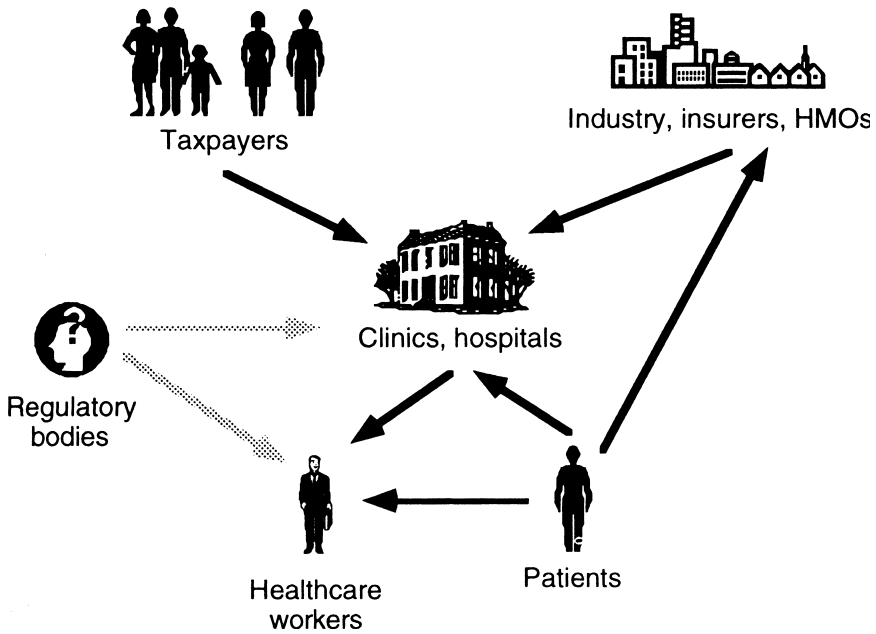

FIGURE 11.1. Some of the actors involved in health care delivery, administration, policy making, and regulation, each of whom may have a stake in an evaluation study. (*Source*: Friedman and Wyatt, 1997a.)

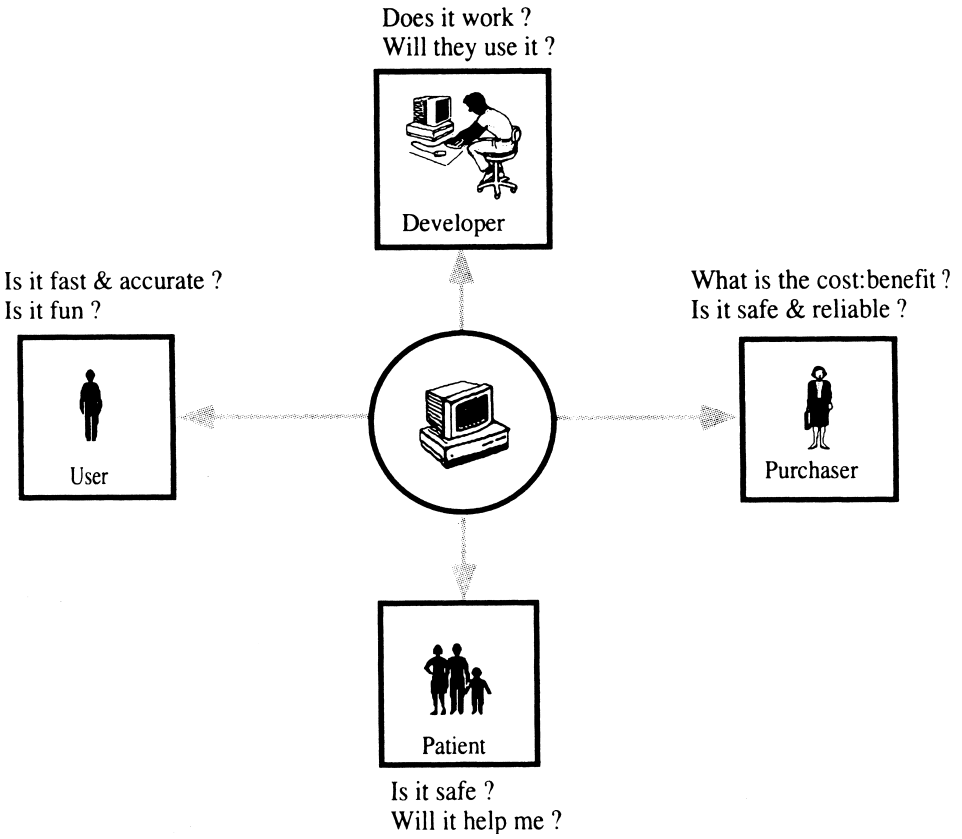

FIGURE 11.2. Different stakeholders may have quite different perspectives on a clinical information resource and questions that they wish to be answered by an evaluation study. (*Source*: Friedman and Wyatt, 1997a.)

(2) computer-based information systems, and (3) the general methodology of study conduct itself. Because of the complexity of each area, any work that combines them necessarily poses serious challenges.

11.2.1 The Complexity of Medicine and Health Care Delivery

Donabedian (1996) informs us that any health care innovation may influence three aspects of the health care system. The first is the health care system's **structure**, including the space it takes up; the equipment available; the financial resources required; and the number, skills, and interrelationships of the staff. The second is the **processes** that take place during health care activity, such as the number and appropriateness of diagnoses, and the investigations and therapies administered. The third is the health

care **outcomes** for both individual patients and the community, such as quality of life, complications of procedures, and length of survival. Thus, when we study the influence of an information resource on a health care system, we may see effects on any of these three aspects. An information resource may lead to an improvement in one area (e.g., patient outcomes) but to deterioration in another (e.g., the costs of running the service).

Also, it is well known that the roles of nursing and clinical personnel are well defined and hierarchical in comparison to those in many other professions. Thus information resources designed for one specific group of professionals, such as a residents' information system designed for one hospital (Young, 1980), may be of little benefit to other groups.

Because health care is a safety-critical area, with more limited budgets and a less tangible currency than, for example, retail or manufacturing, rigorous proof of safety and effectiveness is required in evaluation studies of clinical information resources. Complex regulations apply to people who develop or market clinical therapies or investigational technology. It is not yet clear whether these regulations apply to all computer-based information resources or to only those that manage patients directly, without a human intermediary (Brannigan, 1991).

Medicine is well known to be a complex domain. Students spend a minimum of 7 years gaining qualifications. A single internal-medicine textbook contains approximately 600,000 facts (Wyatt, 1991b); practicing experts have as many as 2 to 5 million facts at their fingertips (Pauker et al., 1976). Also, medical knowledge itself (Wyatt, 1991b), and methods of health care delivery, change rapidly so that the goalposts for a medical information resource may move during the course of an evaluation study.

Patients often suffer from multiple diseases, which may evolve over time at differing rates and may be subject to a number of interventions and other influences over the course of the study period, confounding the effects of changes in information management. There is even variation in how doctors interpret patient data (e.g., prostate-specific antigen results) across medical centers. Thus, simply because an information resource is safe and effective when used in one center on patients who have a given diagnosis, we are not entitled to prejudge the results of using it in another center or with patients who have a different disease profile.

The causal links between introducing an information resource and achieving improvements in patient outcome are long and complex compared with those for direct patient care interventions such as medications. In addition, the functioning and influence of an information resource may depend critically on input from health care workers or patients. It is thus unrealistic to look for quantifiable changes in patient outcomes after the introduction of many information resources until we have documented changes in the structure or processes of health care delivery.

The processes of medical decision making are complex and have been studied extensively (Elstein et al., 1978; Patel. et. al, 2001). Clinicians make many kinds of decisions—including diagnosis, monitoring, therapy, and prognosis—using incomplete and fuzzy data, some of which are appreciated intuitively and are not recorded in the clinical notes. If an information resource generates more effective management of both patient data and medical knowledge, it may intervene in the process of medical decision

making in a number of ways, making difficult the determination of which component of the resource is responsible for observed changes.

There is a general lack of gold standards in medicine. Thus, for example, diagnoses are rarely known with 100 percent certainty, because it is unethical to do all possible tests on every patient, (to follow up patients without good cause), because tests and ability to interpret them are imperfect, and because the human body is simply too complex. When a clinician attempts to establish a diagnosis or the cause of death, even if it is possible to perform a postmortem examination, correlating the patients' symptoms or clinical findings before death with the observed changes may prove impossible. Determining the correct management for a patient is even more complicated, because there is wide variation in **consensus opinions** (Leitch, 1989), as reflected in wide variations in clinical practice even in neighboring areas.

Doctors practice under strict legal and ethical obligations to give their patients the best care that is available, to do patients no harm, to keep patients informed about the risks of all procedures and therapies, and to maintain confidentiality. These obligations may well impinge on the design of evaluation studies. For example, because health care workers have imperfect memories and patients take holidays and participate in the unpredictable activities of real life, it is impossible to impose strict discipline in data recording, and study data are often incomplete. Similarly, before a randomized controlled trial can be undertaken, health care workers and patients are entitled to a full explanation of the possible benefits and disadvantages of being allocated to the control and intervention groups before giving their consent.

11.2.2 The Complexity of Computer-Based Information Resources

From the perspective of a computer scientist, the goal of evaluating a computer-based information resource might be to predict that resource's function and effects from a knowledge of its structure. Although software engineering and formal methods for specifying, coding, and evaluating computer programs have become more sophisticated, even systems of modest complexity challenge these techniques. To formally verify a program rigorously (to obtain proof that it performs all and only those functions specified), we must invest effort that increases exponentially with the program's size—the problem is "**NP hard**." Put simply, to test a program rigorously requires the application of every combination of possible input data in all possible orders. Thus, it entails at least n factorial experiments, where n is the number of input data items. The size of n factorial increases exponentially with small increases in n, so the task rapidly becomes unfeasible. In some technology-led projects, the goals of the new information resources are not defined precisely. Developers may be attracted by technology and may produce applications without first demonstrating the existence of a clinical problem that the application is designed to meet (Heathfield and Wyatt, 1993). An example was a conference entitled "Medicine Meets Virtual Reality: Discovering Applications for 3D Multimedia." The [Q2] lack of a clear need for an information resource makes it hard to evaluate the ability of the information resource to alleviate a clinical problem. Although one can still evaluate

the structure and function of the system in isolation, it will be hard to interpret the results of such an evaluation in clinical terms.

Some computer-based systems are able to adapt themselves to their users or to data already acquired, or they may be deliberately tailored to a given institution; it may then be difficult to compare the results of one evaluation with a study of the same information resource conducted at a different time or in another location. Also, the notoriously rapid evolution of computer hardware and software means that the time course of an evaluation study may be greater than the lifetime of the information resource itself.

Medical information resources often contain several distinct components, including the interface, database, reasoning and maintenance programs, patient data, static medical knowledge, and dynamic inferences about the patient, the user, and the current activity of the user. Such information resources may perform a wide range of functions for users. Thus, if evaluators are to answer questions such as, "What part of the information resource is responsible for the observed effect?" or "Why did the information resource fail?" they must be familiar with each component of the information resource, know its functions, and understand potential interactions (Wyatt, 1989, 1991a).

11.2.3 The Complexity of Study Methods

Studies do not focus solely on the structure and function of an information resource, they also address the resource's effects on the care providers who are customarily its users and on patient outcomes. To understand users' actions, investigators must confront the gulf between peoples' private opinions, public statements, and actual behavior. Humans vary widely in their responses to stimuli, both from minute to minute and from one to another, making the results of measurements subject to random and systematic errors. Thus, studies of medical information resources require analytical tools from the behavioral and social sciences, statistics, and other fields.

Studies require test material, such as clinical cases, and information resource users, such as physicians or nurses. Both are often in shorter supply than the study design requires; the availability of patients also is usually overestimated, sometimes many times over. In addition, it may be unclear what kind of cases or users should be recruited for a study. Often, study designers are faced with a trade-off between selecting cases, users, and study settings with high fidelity to real life and selecting those who will help to achieve adequate experimental control. Finally, one of the more important determinants of the results of an evaluation study is the manner in which case data are abstracted and presented to users. For example, we would expect differing results in a study of an information resource's accuracy depending on whether the test data were abstracted by the developers or by the intended users.

There are many reasons for performing studies, ranging from assessing a student's work to formulating health policy to understanding a specific technical advance. Such reasons will in turn determine the kinds of questions that will be asked about the information resource. To help those who are trying to determine the broad goals of an evaluation study, in Table 11.1 we list some of the many questions that can arise about information resources and about their influence on users, patients, and the health care system.

TABLE 11.1. Possible questions that may arise during the study of a medical information resource.

About the resource itself	About the resource's impact
Is there a clinical need for it?	Do people use it?
Does it work?	Do people like it?
Is it reliable?	Does it improve users' efficiency?
Is it accurate?	Does it influence the collection of data?
Is it fast enough?	Does it influence users' decisions?
Is data entry reliable?	For how long do the observed effects last?
Are people likely to use it?	Does it influence users' knowledge or skills?
Which parts cause the effects?	Does it help patients?
How can it be maintained?	Does it change consumption of resources?
How can it be improved?	What might ensue from widespread use?

(*Source*: Friedman and Wyatt, 1997a.)

11.3 The Full Range of What Can Be Studied

When evaluating a medical information resource, there are five major aspects of interest: (1) the clinical need the resource is intended to address, (2) the process used to develop the resource, (3) the resource's intrinsic structure, (4) the functions that the resource carries out, and (5) the resource's effects on users, patients, and other aspects of the clinical environment. In a theoretically complete evaluation, separate studies of a particular resource might address each aspect. In the real world, however, it is difficult to be comprehensive. Over the course of its development and deployment, a resource may be studied many times with the studies in their totality touching on many or most of these aspects, but few resources will be studied completely and many will, inevitably, be studied only minimally.

The evaluation focus changes as we study the different aspects:

1. *The need for the resource*: Evaluators study the clinical status quo absent the resource. They determine the nature of the problems that the resource is intended to address and the frequency with which these problems arise.
2. *The development process*: Evaluators study the skills of the development team and the methodologies employed to understand whether the design is likely to be sound.
3. *The resource's intrinsic structure*: Evaluators study specifications, flowcharts, program codes, and other representations of the resource that they can inspect without running the program.
4. *The resource's functions*: Evaluators study how the resource performs when it is used.
5. *The resource's effects*: Evaluators study not the resource itself but rather its influence on users, patients, and health care organizations.

Several factors characterize an evaluation study:

- *The focus of study*: The focus can be the status quo before introduction of the information resource, the design process adopted, the resource's structure or function, the resource users' simulated decisions or real decisions, or the clinical actions and patient outcomes once the resource is made available in the workplace.

- *Study setting*: Studies of the design process, the resource's structure, and the resource's functions can be conducted outside the active clinical environment, in a laboratory setting, which is easier logistically and may allow greater control over the evaluation process. Studies to elucidate the need for a resource and studies of the resource's effects on users both usually take place in clinical settings. The effects of a resource on patients and health care organizations can take place in only a true clinical setting where the resource is available for use at the time and place where patient-management decisions are made.
- *Clinical data employed*: For many studies, the resource will actually be run. That will require clinical data, which can be simulated data, data abstracted from real patients' records, or actual patient data. Clearly, the kind of data employed in a study has serious implications for the study results and the conclusions that can be drawn.
- *User of the resource*: Most information resources function in interaction with one or more users. In any particular study, the users of the resource can be members of the development team or the evaluation team, or other individuals not representative of those people who will interact with the resource after it is deployed; or the users in a study could be representative of the end users for whose use the resource is ultimately designed. Again, the selection of resource users can affect study results profoundly.
- *The decisions affected by use of the resource*: Many information resources, by providing information or advice to clinicians, seek to influence the decisions made by these clinicians. As a study moves from the laboratory to the clinical setting, the information provided by the resource potentially has greater implications for the decisions being made. Depending on a study's design and purposes, only simulated decisions may be affected (clinicians are asked what they would do, but no action is taken), or real decisions involved in the care of actual patients may be affected.

Table 11.2 lists eight broad types of studies of clinical information resources that can be conducted: the focus of each type, the setting in which it occurs, the kind of clinical data employed as input to the resource, the person who uses the resource during the study, and the kind of clinical decisions affected by the resource during the study. For example, a laboratory-user impact study would be conducted outside the active clinical environment based on simulated or abstracted clinical data. Although it would involve individuals representative of the end-user population, the study would yield primary results derived from simulated clinical decisions, so the clinical care of patients would not be affected. Read across each row of the table to obtain a feel for the contrasts among these study types.

11.4 Approaches to Study Design

Having established a large number of reasons why it can be difficult to study medical information resources, we now introduce the methods that have been developed to address these challenges. We begin by describing a generic structure that all studies share. Then we introduce, in turn, more specific methods of evaluation and the closely related methods of technology assessment.

TABLE 11.2. Generic types of evaluation studies of clinical information resources.

Type of study	Focus of study	Study setting	Kind of patient data	User of resource	Clinical decisions affected by use of the resource
Need validation	The status quo	Field	Real data	None	None
Design validation	Resource design process	Laboratory	None	None	None
Structure validation	Resource structure	Laboratory	None	None	None
Laboratory function	Resource function	Laboratory	Simulated or abstracted data	Developer evaluator, or clinician	None
Field function	Resource function	Field	Real data	Developer evaluator, or clinician	None
Laboratory user impact	Simulated decisions	Laboratory	Simulated or abstracted data	Clinician	Clinicians' simulated decisions
Field user impact	Simulated decisions	Field	Real data	Clinician	Clinicians' simulated decisions
Clinical impact	Patient care and outcomes	Field	Real data	Clinician	Clinicians' real decisions

(*Source*: Friedman and Wyatt, 1997a.)

11.4.1 The Anatomy of All Studies

The structural elements that all studies share are illustrated in Figure 11.3. Evaluations are guided by someone's or some group's need to know. No matter who that someone is—the development team, the funding agency, or other individuals and groups—the evaluation must begin with a process of negotiation to identify the questions that will be a starting point for the study. The outcomes of these negotiations are an understanding of how the evaluation is to be conducted, usually stated in a written contract or agreement, and an initial expression of the questions the evaluation seeks to answer. The next element of the study is investigation: the collection of data to address these questions and, depending on the approach selected, possibly other questions that arise during the study. The mechanisms are numerous, ranging from the performance of the resource on a series of benchmark tasks to observations of users working with the resource.

The next element is a mechanism for reporting the information back to the individuals who need to know it. The format of the report must be in line with the stipulations of the contract; the content of the report follows from the questions asked and the data collected. The report is most often a written document, but it does not have to be—the purposes of some evaluations are well served by oral reports or by live demonstrations. We emphasize that it is the evaluator's obligation to establish a process through which the results of her study are communicated, thus creating the potential for the study's findings to be put to constructive use. No investigator can guarantee a constructive outcome for a study, but there is much they can do to increase the likelihood of a salutary result. Also note that a salutary result of a study is not necessarily one that casts the resource under study in a positive light. A salutary result is one where the stakeholders learn important information from the study findings.

11.4.2 Philosophical Bases of Approaches to Evaluation

Several authors have developed classifications, or **typologie**s, of evaluation methods or approaches. Among the best is that developed in 1980 by Ernest House. A major

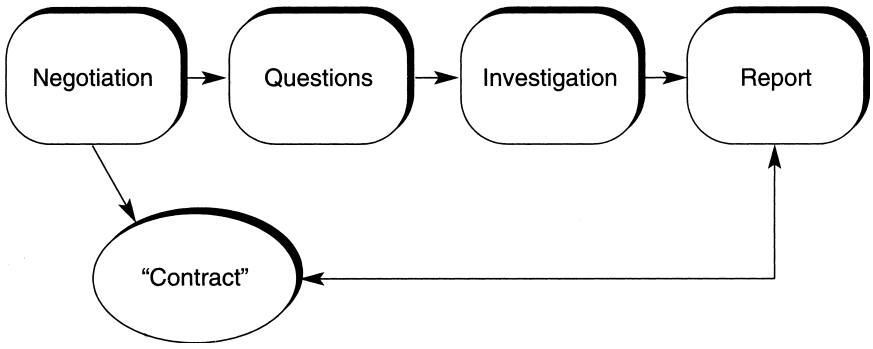

FIGURE 11.3. Anatomy of all evaluation studies. (*Source*: Friedman and Wyatt, 1997a.)

advantage of House's typology is that each approach is linked elegantly to an underlying philosophical model, as detailed in his book. This classification divides current practice into eight discrete approaches, four of which may be viewed as **objectivist** and four of which may be viewed as **subjectivist**. This distinction is very important. Note that these approaches are not entitled objective and subjective, because those words carry strong and fundamentally misleading connotations: of scientific precision in the former case and of imprecise intellectual voyeurism in the latter.

The objectivist approaches derive from a **logical–positivist** philosophical orientation—the same orientation that underlies the classic experimental sciences. The major premises underlying the objectivist approaches are as follows:

- In general, attributes of interest are properties of the resource under study. More specifically, this position suggests that the merit and worth of an information resource—the attributes of most interest in evaluation—can in principle be measured with all observations yielding the same result. It also assumes that an investigator can measure these attributes without affecting how the resource under study functions or is used.
- Rational persons can and should agree on what attributes of a resource are important to measure and what results of these measurements would be identified as a most desirable, correct, or positive outcome. In medical informatics, making this assertion is tantamount to stating that a gold standard of resource performance can always be identified and that all rational individuals can be brought to consensus on what this gold standard is.
- Because numerical measurement allows precise statistical analysis of performance over time or performance in comparison with some alternative, numerical measurement is prima facie superior to a verbal description. Verbal, descriptive data (generally known as qualitative data) are useful in only preliminary studies to identify hypotheses for subsequent, precise analysis using quantitative methods.
- Falsification: while it is possible to disprove a well-formulated scientific hypothesis, it is never possible to fully prove one; thus science proceeds by successive disproof of previously plausible hypotheses.
- Through these kinds of comparisons, it is possible to prove beyond reasonable doubt that a resource is or is not superior to what it replaced or to a competing resource.

[JW3]

Contrast these assumptions with a set of assumptions that derives from an **intuitionist–pluralist** philosophical position that spawns a set of subjectivist approaches to evaluation:

- What is observed about a resource depends in fundamental ways on the observer. Different observers of the same phenomenon might legitimately come to different conclusions. Both can be objective in their appraisals even if they do not agree; it is not necessary that one is right and the other wrong.
- Merit and worth must be explored in context. The value of a resource emerges through study of the resource as it functions in a particular patient care or educational environment.

- Individuals and groups can legitimately hold different perspectives on what constitutes the most desirable outcome of introducing a resource into an environment. There is no reason to expect them to agree, and it may be counterproductive to try to lead them to consensus. An important aspect of an evaluation would be to document the ways in which they disagree.
- Verbal description can be highly illuminating. Qualitative data are valuable, in and of themselves, and can lead to conclusions as convincing as those drawn from quantitative data. The value of qualitative data, therefore, goes far beyond that of identifying issues for later "precise" exploration using quantitative methods.
- Evaluation should be viewed as an exercise in argument, rather than as a demonstration, because any study appears equivocal when subjected to serious scrutiny.

The approaches to evaluation that derive from this subjectivist philosophical perspective may seem strange, imprecise, and unscientific when considered for the first time. This perception stems in large part from the widespread acceptance of the objectivist worldview in biomedicine. The importance and utility of subjectivist approaches in evaluation is, however, emerging. Within medical informatics, there is growing support for such approaches (Rothschild et al., 1990; Forsythe and Buchanan, 1992; Anderson et al., 1995). As stated earlier, the evaluation mindset includes methodological eclecticism. It is important for people trained in classic experimental methods at least to understand, and possibly even to embrace, the subjectivist worldview if they are to conduct fully informative evaluation studies.

11.4.3 Multiple Approaches to Evaluation

House (1980) classifies evaluation into eight approaches. Although most evaluation studies conducted in the real world can be unambiguously tied to one of these approaches, the categories are not mutually exclusive. Some studies exhibit properties of several approaches and are thus not cleanly classified. The first four approaches derive from the objectivist position; the second four are subjectivist.

Comparison Based

The **comparison-based approach** employs experiments and quasi-experiments. The information resource under study is compared with a control condition, a placebo, or a contrasting resource. The comparison is based on a relatively small number of **outcome variables** that are assessed in all groups; randomization, controls, and statistical inference are used to argue that the information resource was the cause of any differences observed. Examples of comparison-based studies include McDonald's work on physician reminders (McDonald et al., 1984a) and the studies from Stanford on rule-based systems (Yu et al., 1979a; Hickam et al., 1985a). The 68 controlled trials of medical decision-support systems reviewed by Hunt and co-workers (1998) fall under the comparison-based approach. The Turing test (Turing, 1950) can be seen as a specific model for a comparison-based evaluation.

Objectives Based

The **objectives-based approach** seeks to determine whether a resource meets its designer's objectives. Ideally, such objectives are stated in great detail, so there is little ambiguity in developing procedures to measure their degree of attainment. These studies are comparative only in the sense that the observed performance of the resource is viewed in relation to stated objectives. The concern is whether the resource is performing up to expectations; it is not whether the resource is outperforming what it replaced. The objectives that are the benchmarks for these studies are typically stated at an early stage of resource development. Although clearly suited to laboratory testing of a new resource, this approach can also be applied to testing of an installed resource. Consider the example of a resource to provide advice to emergency-room physicians (Wyatt, 1989). The designers might set as an objective that the system's advice be available within 15 minutes of the time the patient is first seen. An evaluation study that measured the time for this advice to be delivered, and compared that time with this objective, would be objectives based.

Decision Facilitation

In the **decision facilitation approach**, evaluation seeks to resolve issues important to developers and administrators so that these individuals can make decisions about the future of the resource. The questions that are posed are those that the decision makers state, although the people conducting the evaluation may help the decision makers to frame these questions to be amenable to study. The data-collection methods follow from the questions posed. These studies tend to be formative in focus. The results of studies conducted at the early stages of resource development are used to chart the course of further development, which in turn generates new questions for further study. A systematic study of alternative formats for computer-generated advisories, conducted while the resource to generate the advisories is still under development, is a good example of this approach (de Bliek et al., 1988).

Goal Free

In the three approaches described, the evaluation is guided by a set of goals for the information resource or by specific questions that the developers either state or play a profound role in shaping. Any such study will be polarized by these manifest goals and may be more sensitive to anticipated than to unanticipated effects. In the **goal-free approach**, the people conducting the evaluation are purposefully blinded to the intended effects of an information resource and pursue whatever evidence they can gather to enable them to identify all the effects of the resource, intended or not (Scriven, 1973). This approach is rarely applied in practice, but it is useful to individuals designing evaluations to remind them of the many effects an information resource can engender.

Quasi-Legal

The **quasi-legal approach** establishes a mock trial, or other formal adversary proceeding, to judge a resource. Proponents and opponents of the resource offer testimony and may be examined and cross-examined in a manner resembling standard courtroom procedure. A jury that is witness to the proceedings can then, on the basis of this testimony, make a decision about the merit of the resource. As in a debate, the issue can be decided by the persuasive power of rhetoric as well as by the persuasive power of what is portrayed as fact. There are few examples of this technique formally applied to medical informatics, but the technique has been applied to facilitate difficult decisions in other medical areas such as treatment of sickle cell disease (Smith, 1992).

Art Criticism

The **art criticism approach** relies on methods of art criticism and the principle of connoisseurship (Eisner, 1991). Under this approach, an experienced and respected critic, who may or may not be trained in the domain of the resource but who has a great deal of experience with resources of this generic type, works with the resource. The critic then writes a review highlighting the benefits and shortcomings of the resource. Clearly, the art criticism approach cannot be definitive if the critic is not an expert in the subject domain of a medical informatics resource, because the critic will be unable to judge the clinical or scientific accuracy of the resource's knowledge base or of the advice that it provides. Nonetheless, the thoughtful and articulate comments of an experienced reviewer can help other people to appreciate important features of a resource. Software reviews are examples of this approach in common practice.

Professional Review

The **professional-review approach** is well known in the form of **site visits**. This approach employs panels of experienced peers who spend several days in the environment where the resource is installed. Site visits are often guided by a set of guidelines specific to the type of project under study but sufficiently generic to accord the reviewers a great deal of control over the conduct of any particular visit. The reviewers are generally free to speak with whomever they wish and to ask these individuals whatever they consider important to know. They may also request documents for review. Over the course of a site visit, unanticipated issues may emerge. The site visitors typically explore both the anticipated issues and the questions articulated in the guidelines and those that emerge during the site visit itself. The result is a report usually drafted on site or very soon after the visit is completed.

Responsive–Illuminative

The **responsive–illuminative approach** seeks to represent the viewpoints of both users of the resource and people who are an otherwise significant part of the clinical environment where the resource operates (Hamilton et al., 1977). The goal is understanding or

illumination rather than judgment. The methods used derive largely from ethnography. The investigators immerse themselves in the environment where the resource is operational. The designs of these studies are not rigidly predetermined. They develop dynamically as the investigators' experience accumulates. The study team begins with a minimal set of orienting questions; the deeper questions that receive thoroughgoing study evolve over time. Many examples of studies using this approach can be found in the literature of medical informatics (Fafchamps et al., 1991; Forsythe, 1992; Ash et. al., 2003).

Note that the study types described in Table 11.2 relate to the purposes, foci, settings, and logistics of evaluation studies. The evaluation approaches introduced in this section address a complementary issue, What methods will be used to identify specific questions and to collect data as part of the actual conduct of these studies? Although it is perhaps extreme to state that every evaluation approach can apply to every type of study, there is certainly potential to use both objectivist and subjectivist approaches throughout Table 11.2. At the two extremes, for example, both need-validation studies and clinical-effects studies provide opportunities for application of subjectivist as well as objectivist approaches.

11.4.4 Stages of Technology Assessment

Yet another way to categorize studies is according to the three stages of technology assessment (Fuchs and Garber, 1990; Garber and Owens, 1994). The first stage emphasizes **technical characteristics**, such as the response time of an information system to a query or the resolution of an imaging system. The second stage emphasizes the **efficacy** or effectiveness of a device, information system, or diagnostic or therapeutic strategy (Fuchs and Garber, 1990). Clinical trials of information systems usually fit this category, as do randomized trials of clinical interventions. The trials often use **process measures**, such as the degree of physician compliance with computer-generated reminders or the change in laboratory parameters in response to treatment rather than the endpoints that matter to patients: mortality, morbidity, and cost. Studies that determine the sensitivity and specificity of diagnostic tests are another example of second-stage assessments (see Chapter 3).

Third-stage assessments directly evaluate effectiveness via health and economic **outcomes**; therefore, these evaluations are the most comprehensive technology assessments (Fuchs and Garber, 1990). A third-stage evaluation of a computer-based reminder system for breast cancer screening would examine changes in mortality or morbidity from breast cancer rather than physician compliance with guidelines. Typically, a third-stage evaluation also would evaluate the costs of such a system. When outcomes are infrequent or occur after a long delay (such as the occurrence of breast cancer), third-stage evaluations may be substantially more difficult to perform than are second-stage evaluations; thus, third-stage assessments are uncommon in medical informatics (see Hunt et al., 1998). Third-stage evaluations also may consider the importance of patients' preferences in assessing the outcomes of an intervention (Nease and Owens, 1994; Owens, 1998a).

We now examine the types of studies that investigators may initiate for each of the stages of technology assessment.

Stage I Assessments: Technical Characteristics

The choice of what to evaluate during a first-stage technology assessment depends on the purpose of the evaluation (Friedman and Wyatt, 1997b). Possibilities include the evaluation of the design and development process of a clinical information resource or of the structure of the resource (the hardware, input and output devices, user interface, internal data, knowledge structures, processor, algorithms, or inference methods). An assessment of the design and development process could evaluate the software engineering of the resource. Such an evaluation might be important to assess how the resource could be integrated with other systems or platforms. The rationale for studying the structure of the resource is the assumption that, if the resource contains appropriately designed components linked together in a suitable architecture, the system is more likely to function correctly.

Stage II Assessments: Clinical Efficacy

Second-stage assessments move beyond evaluation of operating parameters to an evaluation of the function of the information resource. These evaluations are increasingly common. Recent systematic reviews report over 100 clinical trials of information resources (Balas et al., 1996; Shea et al., 1996; Hunt et al., 1998). Examples of second-stage evaluations include studies of computer-assisted drug dosing; preventive care reminder systems; and computer-aided quality assurance programs for active medical problems. The majority of these second-stage evaluations assess the effect of information resources on the process of care. Did the clinician prescribe the right drug dose? Did the patient receive an influenza vaccine? In situations in which a process measure correlates closely with health outcome (e.g., use of thrombolytic therapy for patients who have heart attacks correlates closely with decreased mortality rates), use of the process measure will not adversely affect validity and will increase the feasibility of the study (Mant and Hicks, 1995). The link from many interventions to the intended health and economic outcomes is not, however, well defined; in these circumstances, a second-stage technology assessment may not be sufficient to justify implementation of a system, particularly if the system is costly.

Stage III Assessments: Comprehensive Clinical Effectiveness, Economic, and Social Outcomes

Rising health care costs have forced policymakers, clinicians, and developers to assess whether health interventions provide sufficient value for the required economic investment. Thus, a demonstration of efficacy is often not sufficient. Proponents of a technology must also establish its cost effectiveness (see Section 11.5.5 for a more detailed explanation of cost-effectiveness studies). The third stage of technology assessment encompasses these more sophisticated assessments. The hallmark of these evaluations is a comprehensive assessment of health and economic outcomes. Studies that evaluate comprehensive outcomes will be more useful than studies that evaluate narrowly defined outcomes. Thus, a study that evaluates the cost

effectiveness of an information resource in terms of dollars per quality-adjusted life year (QALY) saved (see Chapter 3) would enable clinicians and policymakers to compare the cost effectiveness of an information resource to a wide variety of other interventions. In contrast, a study that evaluates an information resource in terms of dollars per case of cancer prevented would provide a useful comparison only for other interventions that prevent cancer.

The choice of outcome measures for a third-stage assessment depends on the purpose of the study and on the cost and feasibility of measuring the outcome. Common choices include the number of lives saved, life-years saved, quality-adjusted life years saved, cancers prevented, and cases of disease averted. For example, a third-stage evaluation of a computer-generated protocol for treatment of hypertension could measure changes in blood pressure of patients whose care was governed by the protocol. The evaluation could also assess the costs of implementing the protocol and subsequently the cost effectiveness of the implementation of the computer-generated protocol. An evaluation of computer-based guidelines for care of people who have human immunodeficiency virus (HIV) evaluated the effect of the guideline on the rate of hospitalization for opportunistic infection (Safran et al., 1995). The study found that, under the guidelines, providers responded more rapidly to changes in patient status (such as abnormal laboratory tests), but this prompt action did not change the rate of hospitalization. This study highlights the difficulty of demonstrating that a beneficial change in the process of care has led to improved health outcomes. In fact, few studies have demonstrated that information resources improve health outcomes (Hunt et al., 1998). The studies may not show benefit because of inadequate sample sizes, use of outcome measures that are difficult to assess, inadequate follow-up, other weaknesses in study design, or interventions that do not work (Rotman et al., 1996).

In summary, the requirements of a technology assessment have expanded to include comprehensive health, economic, and social outcomes. Third-stage technology assessment is a particular challenge in medical informatics. Although the use of process measures will be appropriate for third-stage assessment when evidence from systematic reviews shows that process measures correlate very well with patient outcomes, until that time investigators will need to plan studies that explicitly incorporate comprehensive health and economic outcomes.

11.5 Conduct of Objectivist Studies

In this section, we focus on the comparison-based approach, which is the most widely used objectivist approach and which also is the basis of most work in technology assessment.

11.5.1 Structure and Terminology of Comparative Studies

In a comparative study, the investigator typically creates a contrasting set of conditions to compare the effects of one with those of another. Usually, the goal is to attribute cause and effect or to answer scientific questions raised by other kinds of studies. After

identifying a sample of subjects for the study, the researcher assigns each subject, often randomly, to one or a set of conditions. Some variable of interest is measured for each subject. The aggregated values of this variable are compared across the conditions. To understand the many issues that affect design of comparative studies, we must develop a precise terminology.

The **subjects** in a study are the entities about which data are collected. A specific study will employ one sample of subjects, although this sample might be subdivided if, for example, subjects are assigned to conditions in a comparative design. It is key to emphasize that subjects are often people—either care providers or recipients—but also may be information resources, groups of people, or organizations. In informatics, medical care is conducted in hierarchical settings with naturally occurring groups (a "doctor's patients"; the "care providers in a ward team"), so we often face the challenging question of exactly who the subjects are.

Variables are specific characteristics of the subjects that either are measured purposefully by the investigator or are self-evident properties of the subjects that do not require measurement. In the simplest study, there may be only one variable, for
[JW4] example, the time required for a user of an information system to complete a particular task.

Some variables take on a continuous range of values. Others have a discrete set of levels corresponding to each of the measured values that that variable can have. For example, in a hospital setting, physician members of a ward team can be classified as residents, fellows, or attendings. In this case, the variable "physician's level of qualification" has three levels.

The **dependent variables** form a subset of the variables in the study that captures the outcomes of interest to the investigator. For this reason, dependent variables are also called **outcome variables**. A study may have one or more dependent variables. In a typical study, the dependent variable will be computed, for each subject, as an average over a number of tasks. For example, clinicians' diagnostic performance may be measured over a set of cases, or "tasks", that provide a range of diagnostic challenges. (In studies in which computers or people solve problems or work through clinical cases, we use the term "task" generically to refer to those problems or cases. Designing or choosing tasks can be the most challenging aspect of an evaluation.)

The **independent variables** are included in a study to explain the measured values of the dependent variables. For example, whether a computer system is available, or not, to support certain clinical tasks could be the major independent variable in a study designed to evaluate that system. A purely descriptive study has no independent variables; comparative studies can have one or many independent variables.

Measurement challenges almost always arise in the assessment of the outcome or dependent variable for a study. Often, for example, the dependent variable is some type of performance measure that invokes concerns about reliability (precision) and validity (accuracy) of measurement. Depending on the study, the independent variables may also raise measurement challenges. When the independent variable is gender, for example, the measurement problems are relatively straightforward. If the independent variable is an attitude, level of experience, or extent of resource use, however, profound measurement challenges can arise.

11.5.2 Issues of Measurement

Measurement is the process of assigning a value corresponding to the presence, absence, or degree of a specific attribute in a specific object, as illustrated in Figure 11.4. Measurement usually results in either (1) the assignment of a numerical score representing the extent to which the attribute of interest is present in the object, or (2) the assignment of an object to a specific category. Taking the temperature (attribute) of a patient (object) is an example of the process of measurement.[2]

From the premises underlying objectivist studies (see Section 11.4.2), it follows that proper execution of such studies requires careful and specific attention to methods of measurement. It can never be assumed, particularly in informatics, that attributes of interest are measured without error. Accurate and precise measurement must not be an afterthought. Measurement is of particular importance in medical informatics because, as a relatively young field, informatics does not have a well-established tradition of "variables worth measuring" or proven instruments for measuring them. By and large, people planning studies are faced first with the task of deciding what to measure and then with that of developing their own measurement methods. For most researchers, these tasks prove to be harder and more time consuming than initially anticipated. In some cases, informatics investigators can adapt the measures used by other investigators, but often they need to apply their measures to a different setting where prior experience may not apply.

We can underscore the importance of measurement by establishing a formal distinction between studies undertaken to develop methods for making measurements, which

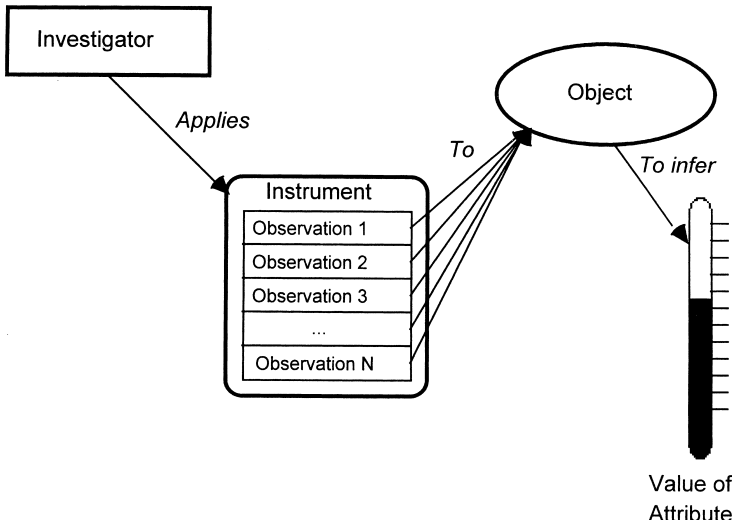

FIGURE 11.4. The process of measurement. (*Source*: Friedman and Wyatt, 1997a.)

[2]When we speak specifically of measurement, it is customary to use the term "object" to refer to the entity on which measurements are made.

we call measurement studies, and the subsequent use of these methods to address questions of direct importance in informatics, which we call demonstration studies. **Measurement studies** seek to determine how accurately an attribute of interest can be measured in a population of objects. In an ideal objectivist measurement, all observers will agree on the result of the measurement. Any disagreement is therefore due to error, which should be minimized. The more agreement among observers or across observations, the better the measurement. Measurement procedures developed and validated through measurement studies provide researchers with what they need to conduct **demonstration studies** that directly address questions of substantive and practical concern. Once we know how accurately we can measure an attribute using a particular procedure, we can employ the measured values of this attribute as a variable in a demonstration study to draw inferences about the performance, perceptions, or effects of an information resource. For example, once a measurement study has explored how accurately the speed of an information resource can be measured, a related demonstration study would explore whether a particular resource has sufficient speed—speed being measured using methods developed in the measurement study—to meet the needs of busy clinicians.

A detailed discussion of measurement issues is beyond the scope of this chapter. The bottom line is that investigators should know that their measurement methods will be adequate before they collect data for their studies. It is necessary to perform a measurement study, involving data collection on a small scale, to establish the adequacy of all measurement procedures if the measures to be used do not have an established track record. Even if the measurement procedures of interest do have a track record in a particular health care environment and with a specific mix of cases and care providers, they may not perform equally well in a different environment, so measurement studies may still be necessary. Researchers should always ask themselves, "How good are my measures in this particular setting?" whenever they are planning a study, before they proceed to the demonstration phase. The importance of measurement studies for informatics was explained in 1990 by Michaelis and co-workers. A more recent study (Friedman et al., 2003) has demonstrated that outcome studies of clinical information systems have not systematically addressed the adequacy of the methods used to measure the specific outcomes reported in these studies.

Whenever possible, investigators planning studies should employ established measurement methods, with a "track record", rather than developing their own. While there exist relatively few compendia and measurement instruments specifically for informatics, one such list is provided in Friedman and Wyatt (1997a). A Web-based resource listing over 50 instruments asscociated with the development, usability, and impact of information systems at http://www.isworld.org/ surveyinstruments/surveyinstruments.htm.

11.5.3 Control Strategies in Comparative Studies

One of the most challenging questions in comparative study design is how to obtain control. We need a way to monitor all the other changes taking place that are not attributable to the information resource. In clinical medicine, it is occasionally possible to predict patient outcomes with good accuracy from a small set of initial clinical find-

ings—for example, the survival of patients in intensive care (Knaus et al., 1991). In these unusual circumstances, where we have a mechanism to tell us what would have happened to patients if we had not intervened, we can compare what actually happens with what is predicted to draw tentative conclusions about the benefit of the information resource. Such accurate predictive models are, however, extremely unusual in medicine (Wyatt and Altman, 1995). Instead, we use various types of controls: subjects who complete tasks that are not affected by the intervention of interest.

In the following sections we review a series of control strategies. We employ as a running example a reminder system that prompts doctors to order prophylactic antibiotics for orthopedic patients to prevent postoperative infections. In this example, the intervention is the installation and commissioning of the reminder system; the subjects are the physicians; and the tasks are the patients cared for by the physicians. The dependent variables derive from the outcome measurements made and would include physicians' ordering of antibiotics and the rate of postoperative infections averaged across the patients cared for by each physician.

Descriptive (Uncontrolled) Studies

In the simplest possible design, an uncontrolled or **descriptive study**, we install the reminder system, allow a suitable period for training, and then make our measurements. There is no independent variable. Suppose that we discover that the overall postoperative infection rate is 5 percent and that physicians order prophylactic antibiotics in 60 percent of orthopedic cases. Although we have two measured dependent variables, it is hard to interpret these figures without any comparison—it is possible that there has been no change due to the system.

Historically Controlled Experiments

As a first improvement to a descriptive study, let us consider a **historically controlled experiment**, sometimes called a **before–after study**. The investigator makes **baseline** measurements of antibiotic ordering and postoperative infection rates before the information resource is installed, and then makes the same measurements after the information resource is in routine use. The independent variable is time and has two levels: before and after resource installation. Let us say that, at baseline, the postoperative infection rates were 10 percent and doctors ordered prophylactic antibiotics in only 40 percent of cases; the postintervention figures are the same as before (see Table 11.3).

The evaluators may claim that the halving of the infection rate can be safely ascribed to the information resource, especially because it was accompanied by a 20 percent improvement in doctors' antibiotic prescribing. Many other factors might, however, have changed in the interim to cause these results, especially if there was a long interval between the baseline and postintervention measurements. New staff could have taken over, the case mix of patients could have altered, new prophylactic antibiotics may have been introduced, or clinical audit meetings may have highlighted the infection problem and thus caused greater clinical awareness. Simply assuming that the reminder system alone caused the reduction in infection rates is naive. Other factors, known or unknown,

TABLE 11.3. Hypothetical results of historically controlled study of an antibiotic reminder system.

	Antibiotic prescribing rate	Postoperative infection rate
Baseline results (before installation)	40%	10%
Postinstallation results	60%	5%

(*Source*: Friedman and Wyatt, 1997a.)

could have changed meanwhile, making untenable the simple assumption that our intervention is responsible for all of the observed effects.

An improvement on this design is to add either internal or external controls—preferably both. The internal control should be a measure likely to be affected by any nonspecific changes happening in the local environment, but unaffected by the intervention. The external control can be exactly the same measure as in the target environment, but in a similar external setting, e.g., another hospital. If the measure of interest changes while there is no change in either internal or external controls, a skeptic needs to be quite resourceful to claim that the system is not responsible

Simultaneous Nonrandomized Controls

To address some of the problems with historical controls, we might use **simultaneous controls**, which requires us to make our outcome measurements in doctors and patients who are not influenced by the prophylactic antibiotic reminder system but who are subject to the other changes taking place in the environment. Taking measurements both before and during the intervention strengthens the design, because it gives an estimate of the changes due to the nonspecific factors taking place during the study period.

This study design would be a parallel group comparative study with simultaneous controls. Table 11.4 gives hypothetical results of such a study, focusing on postoperative infection rates as a single outcome measure or dependent variable. The independent variables are time and group, both of which have two levels of intervention and control. There is the same improvement in the group where reminders were available, but no improvement—indeed a slight deterioration—where no reminders were available. This design provides suggestive evidence of an improvement that is most likely to be due to the reminder system. This inference is stronger if the same doctors worked in the same wards during the period the system was introduced, and if similar kinds of patients, subject to the same nonspecific influences, were being operated on during the whole time period.

TABLE 11.4. Hypothetical results of simultaneous controlled study of antibiotic reminder system.

	Postoperative infection rates	
	Reminder group	Control group
Baseline results	10%	10%
Postintervention results	5%	11%

(*Source*: Friedman and Wyatt, 1997a.)

Even though the controls in this example are simultaneous, skeptics may still refute our argument by claiming that there is some systematic, unknown difference between the clinicians or patients in the two groups. For example, if the two groups comprised the patients and clinicians in two adjacent wards, the difference in the infection rates could be attributable to systematic or chance differences between the wards. Perhaps hospital-staffing levels improved in some wards but not in others, or there was cross infection by a multiple-resistant organism only among the patients in the control ward. To overcome such criticisms, we could expand the study to include all wards in the hospital—or even other hospitals—but that would clearly take considerable resources. We could try to measure everything that happens to every patient in both wards and to build complete psychological profiles of all staff to rule out systematic differences. We would still, however, be vulnerable to the accusation that some variable that we did not measure—did not even know about—explains the difference between the two wards. A better strategy is to ensure that the controls really are comparable by randomizing them.

Simultaneous Randomized Controls

The crucial problem in the previous example is that, although the controls were simultaneous, there may have been systematic, unmeasured differences between them and the subjects receiving the intervention. A simple and effective way of removing systematic differences, whether due to known or unknown factors, is to randomize the assignment of subjects to control or intervention groups. Thus, we could randomly allocate one-half of the doctors on both wards to receive the antibiotic reminders and the remaining doctors to work normally. We would then measure and compare postoperative infection rates in patients managed by doctors in the reminder and control groups. Provided that the doctors never look after one another's patients, any difference that is statistically "significant" (conventionally, for which the p value is less than 0.05) can be attributed reliably to the reminders. The only way other differences could have emerged is by chance.

Table 11.5 shows the hypothetical results of such a study. The baseline infection rates in the patients managed by the two groups of doctors are similar, as we would expect, because the patients were allocated to the groups by chance. There is a greater reduction in infection rates in patients of reminder physicians compared with those of control physicians. Because random assignment means that there was no systematic difference in patient characteristics between groups, the only systematic difference between the two groups of patients is receipt of reminders by their doctors.

Table 11.5. Hypothetical results of a simultaneous randomized controlled study of antibiotic reminder system.

	Postoperative infection rates	
	Reminder physicians	Control physicians
Baseline results	11%	10%
Postinstallation results	6%	8%

(*Source*: Friedman and Wyatt, 1997a.)

Provided that the sample size is large enough for these results to be statistically significant, we might begin to conclude with some confidence that providing doctors with reminders caused the reduction in infection rates. One lingering question is why there was also a small reduction, from baseline to installation, in infection rates in control cases, even though the control group should have received no reminders (Wyatt and Wyatt, 2003).

11.5.4 Threats to Inference and Validity

We all want our studies to be valid. There are two aspects to this: internal and external validity. If a study has **internal validity**, we can be confident in the conclusions drawn from the specific circumstances of the experiment—the population of subjects studied, the measurements made, and the interventions provided. Are we justified in concluding that the differences observed are due to the attributed causes? Even if all threats to internal validity are overcome to our satisfaction, we would also like our study to have **external validity**, such that the conclusions can be generalized from the specific setting, subjects, and intervention studied to the broader range of settings that other people will encounter. Thus, even if we demonstrate convincingly that our antibiotic reminder system reduces postoperative infection rates in our own hospital, this finding is of little interest to other clinicians unless we can convince them that the results can be generalized safely to other reminder systems or to the same system in other hospitals.

When we conduct a comparative study, there are four possible outcomes. We illustrate them in the context of a study that explores the effectiveness of an information resource and that uses an appropriate comparative design as follows:

1. The information resource was truly effective, and our study shows that it was.
2. The information resource was truly ineffective, and our study shows that it was.
3. The information resource was truly effective, but for some reason our study mistakenly failed to show that it was.
4. The information resource was truly ineffective, but for some reason our study mistakenly suggested that it was effective.

Outcomes 1 and 2 are salutary from a methodological viewpoint, the results of the study mirror reality. Outcome 3 is a false-negative result, or **type II error**. In the language of inferential statistics, we mistakenly accept the **null hypothesis**. Type II errors can arise because the size of the information resource's effect on the measure of interest is small and too few subjects have been included for the study to detect it (Freiman et al., 1978). Alternatively, we may have failed to measure the outcome variable on which the example resource is having an effect. In outcome 4, we have concluded that the resource is valuable when it is not; we have a false-positive result or **type I error**. We have mistakenly rejected the null hypothesis. A risk of a type I error is built into every study. When we accept, for example, the conventional value of p, 0.05 as a criterion for statistical significance, we are consciously accepting a 5 percent risk of making a type I error as a consequence of using randomization as a mechanism of experimental control. If we feel uncomfortable with this 5 percent risk of a false-positive result or type I error, we can

reduce it by reducing the threshold for statistical significance to 0.01, which carries only a 1 percent chance of a type I error.

The more important threats to internal validity of studies are the following:

- **Assessment bias**: It is important to ensure that all persons involved in making measurements do not allow their own feelings and beliefs about an information resource—positive or negative—to bias the results. Consider a study in which the same clinicians who are users of an antibiotic reminder system also collect the clinical data used for determining whether the advice generated by the system is correct, such as the incidence of significant wound or chest infections. If they had some prejudice against the reminder system and wished to undermine it, they might massage the clinical infection data to prove themselves right and the reminder system wrong in certain patients. Thus, they might record that a patient was suffering from a nonexistent postoperative cough with productive sputum to justify an antibiotic prescription that the reminder system had not advised.
- **Allocation bias**: Early studies of information resources often take place in the environment in which the resources were developed and often arouse strong (positive or negative) feelings among study subjects. In a study where patients are randomized and the subjects have strong beliefs about the information resource, two biases may arise. Investigators may cheat the randomization method and systematically allocate easier (or more difficult) cases to the information resource group (allocation bias), or they may avoid recruiting a particularly easy (or difficult) case to the study if they know in advance that the next patient will be allocated to the control group (Schulz et al., 1995).
- The **Hawthorne effect**: The Hawthorne effect is the tendency for humans to improve their performance if they know it is being studied. Psychologists measured the effect of ambient lighting on workers' productivity at the Hawthorne factory in Chicago (Roethligsburger and Dickson, 1939). Productivity increased as the room illumination level was raised, but productivity increased again when the illumination level was accidentally reduced. The study itself, rather than changes in illumination, caused the increases. During a study of a medical information resource, the attention of the investigators can lead to an improvement in the performance of all subjects in all study groups, intervention and control, due to the Hawthorne effect.
- The **checklist effect**: The checklist effect is the improvement observed in decision making due to more complete and better-structured data collection when paper-based or computer-based forms are used to collect patient data. The effect of forms on decision making can equal that of computer-generated advice (Adams et al., 1986), so it must either be controlled for or quantified. To control for the checklist effect, investigators should collect the same data in the same way in the control and information resource groups, even though the information resource's output is available only in the latter group (see for e.g., Wyatt, 1989).
- The **placebo effect**: In some drug trials, simply giving patients an inactive tablet or other placebo can cause a measurable improvement in some clinical variables such as well-being, sleep pattern, and exercise tolerance, because patients feel good about receiving attention and potentially useful medication. This placebo effect may be

more powerful than the drug effect itself and may obscure a complete absence of pharmaceutical benefit. In a study of a medical information resource, if some patients watch their doctors consult an impressive workstation while others have no such experience, this experience could unbalance the groups and overestimate the value of the information resource. Alternatively, some patients might believe that a care provider who needs a computer workstation is less competent than one who can manage without. Several studies have, however, shown that patients show more confidence in clinicians who use technology than in those who get by without it.

These and a number of other biases that may apply in certain kinds of studies are more fully discussed in Chapter 10 of Friedman and Wyatt (1997a).

11.5.5 Cost-Effectiveness and Cost-Benefit Studies

The purpose of cost-effectiveness and cost-benefit analyses is to assess quantitatively the benefits obtained from a health intervention relative to the costs of the intervention. In short, cost-effectiveness and cost-benefit analyses provide a mechanism to assess the relative value of different interventions in producing health benefits, such as longer life or greater quality of life. Our description of these analyses is brief; for further details, see Gold (1996), Weinstein and Fineberg (1980), and Sox et al. (1988). In a **cost-effectiveness analysis**, the analyst expresses the health benefits in units of health outcomes (e.g., lives saved) and the costs in dollars. The analysts can choose the health outcomes that they believe as appropriate for the purpose of the analysis, such as life years saved, quality-adjusted life years saved, or cases of disease prevented. Usually the analyst seeks to compare the cost and health effects of one treatment relative to the costs and health effects of another treatment. In such a situation, the appropriate estimate of the relative value of the interventions is the **incremental cost-effectiveness ratio**. To calculate the incremental cost-effectiveness ratio for intervention "b" relative to intervention "a", we divide the difference in the costs with the two interventions by the difference in health benefits. For example, for interventions whose benefits are measured as increases in life expectancy (LE), we calculate the incremental cost-effectiveness ratio for intervention "b" relative to intervention "a" as

$$(C_b - C_a)/(LE_b - LE_a)$$

where C_b is the cost of intervention "b", C_a is the cost of intervention "a", LE_b is the life expectancy with intervention "b", and LE_a is life expectancy with intervention "a."

In contrast to cost-effectiveness analyses, a **cost-benefit analysis** values all benefits and costs in dollars. Thus, if a health intervention averts a death, the analyst must express the value of that averted death in dollars. The goal of such an analysis is to determine whether the benefit (expressed in dollars) is larger than the cost (expressed in dollars).

To perform a cost-effectiveness analysis, the analyst must perform the following steps (Office of Technology Assessment, 1980; Gold et al., 1996): (1) define the problem, including identification of the objective and perspective of the analysis, and the alternative interventions under consideration; (2) identify and analyze the benefits;

(3) identify and analyze the costs; (4) perform discounting; (5) analyze uncertainties; (6) address ethical questions; and (7) interpret the results. We shall illustrate these concepts with the example that we used in Section 11.5.3 of a computer-based reminder system that prompts physicians to order prophylactic antibiotics before orthopedic surgery.

The first step of a cost-effectiveness analysis is to define the problem clearly. Failure to define the problem carefully can lead to many difficulties. One approach to defining the problem is to determine the decision context. What decision does the analyst, or the consumer of the analysis, need to make? Is the decision whether to implement a computer-based antibiotic reminder system or a manual system? Or, is the decision which computer-based antibiotic reminder system to implement? Or, do the decision makers seek to know whether to implement a computer-based system or to hire a nurse practitioner to check the antibiotic orders? Answers to these questions will enable the analyst to frame the analysis appropriately.

To ensure that the cost-effectiveness ratio that the analyst calculates will be helpful for the decision makers, the analyst should also identify the objective of the study. Is the objective to reduce hospital costs or all costs? A program that reduced hospital costs could do so by shifting certain costs to the outpatient setting. Is that a concern? The analyst should also determine the perspective of the analysis, because the perspective determines whose costs and whose benefits belong in the analysis. For example, if the perspective is that of the hospital, outpatient costs may not matter. If the perspective of the analysis is societal, however, as is typically true, the analyst should include outpatient costs. Finally, the analyst should identify the alternatives that the decision makers will (or should) consider. Rather than install an expensive hospital information system, perhaps the hospital should sign a contract with another hospital to perform all orthopedic surgeries. If the analyst has evaluated the decision context carefully, the important alternatives should be clear.

The next step in a cost-effectiveness analysis is to identify and analyze the health outcomes and costs of the alternative interventions. How should the analyst evaluate the cost effectiveness of the antibiotic reminder system? First, the analyst should decide how to measure the health benefit of the system, and then could quantify the health benefit by assessing the number of postoperative infections before and after implementation of the computer-based system. The units of the cost-effectiveness ratio would therefore be dollars expended per postoperative infection prevented. Such a ratio may be helpful to decision makers, but the decision makers could compare the cost effectiveness of the system only with other interventions that prevent postoperative infections; for example, they could not compare the cost effectiveness of the system with that of a computer-based reminder system for breast cancer screening. To remedy this problem, the analyst could choose a more comprehensive measure of health outcome, such as quality-adjusted life years (see Chapter 3). The analysts would then estimate the number of quality-adjusted life years saved for each postoperative infection prevented. Decision modeling provides one approach to make such estimates (see Chapter 3). Thus, use of quality-adjusted life years would enable policymakers to evaluate the cost effectiveness of the antibiotic reminder system relative to other interventions, but would impose an additional analytic burden on the analyst.

To evaluate the incremental cost effectiveness of the reminder system, the analysts must also estimate the costs with the old system and the costs with the computer-based system. The literature often refers to *direct costs* and *indirect costs*, but these terms are not used consistently. We shall follow the definitions and conventions of Gold and coworkers (1996) and refer to costs as either direct costs or productivity costs. With this approach, certain costs that we classify as direct costs formerly were considered as indirect costs. The **direct costs** include the value of all the goods, services, and other resources that are required to produce an intervention, including resources consumed because of future consequences (intended or unintended) of the intervention (Gold et al., 1996). Direct costs include changes in the use of health care resources, of nonhealth care resources, of informal caregiver time, and of patient time. The direct health care costs include the costs of drugs, tests, procedures, supplies, health care personnel, and facilities. For the antibiotic reminder system, the direct health care costs include the costs of installation, maintenance, personnel, supplies, drugs (antibiotics), and the future cost savings that may result from a reduction in postoperative infection, among others. Direct nonhealth care costs include other services required for the delivery of an intervention, such as patients' transportation costs associated with medical care. If family members provide ancillary care, the value of their time is also a cost of the intervention. The time a patient must spend to receive the intervention is also a cost. Because implementation of an antibiotic reminder system would not change these costs, the analyst does not need to include them in an analysis. **Productivity costs** are those costs that accrue because of changes in productivity due to illness or death; they could be relevant to the analysis of the antibiotic reminder system if prevention of postoperative infection changed substantially the time away from work for patients in whom infection was prevented (for further discussion, see Gold et al., 1996).

To complete the analysis, the analyst should discount health and economic outcomes, address uncertainty and ethical considerations, and interpret the results. **Discounting** enables the analyst to account for time preference in the analysis: Expenditures and health benefits that occur in the future have less value than do those expenditures or benefits that occur immediately. The analyst performs discounting by calculating the net present value of health outcomes and costs; this calculation reduces the influence of future health and economic outcomes relative to those that occur in the present (for further explanations, see Gold et al., 1996). Both health and economic outcomes should be discounted (Gold et al., 1996). Sensitivity analyses (described in Chapter 3) provide a mechanism for assessing the importance of uncertainty. Ethical concerns include how to ensure equity in policy alternatives, how to value outcomes, and how to choose a cost-effectiveness threshold (also see Chapter 10). The **cost-effectiveness threshold** (e.g., $50,000 per quality-adjusted life year saved) reflects the value judgment of the decision makers about the maximum value of a year of life saved. Although currently there is not a consensus on the appropriate threshold, many interventions that are used widely cost less than $50,000 to $60,000 per quality-adjusted life year gained (Owens, 1998b). Interpretation of the results should incorporate statements both about the influence of uncertainty on the estimated cost-effectiveness ratio and about ethical concerns.

Cost-effectiveness and cost-benefit analyses provide tools for helping policymakers and clinicians to understand the relationship between the health outcomes and costs of

alternative health interventions, including information resources. We emphasize that they provide information about one important aspect of an intervention, or information system, but are insufficient alone for decision making. Other social, ethical, and political factors will be important for most decisions. Evaluation of comprehensive information systems poses formidable challenges because the benefits of such systems may be diffuse, varied, and difficult to quantify. Nonetheless, like other health care interventions and innovations, information resources must provide sufficient benefit to justify their expense.

11.6 Conduct of Subjectivist Studies

The objectivist approaches to evaluation, described in the previous section, are useful for addressing some, but not all, of the interesting and important questions that challenge investigators in medical informatics. The subjectivist approaches described here address the problem of evaluation from a different set of premises. They use different but equally rigorous methods.

11.6.1 The Rationale for Subjectivist Studies

Subjectivist methods enable us to address the deeper questions that arise in informatics: the detailed "whys" and "according to whoms" in addition to the aggregate "whethers" and "whats." As defined earlier, the responsive–illuminative approach, within the subjectivist family of approaches, seeks to represent the viewpoints of people who are users of the resource or are otherwise significant participants in the clinical environment where the resource operates. The goal is illumination rather than judgment. The investigators seek to build an argument that promotes deeper understanding of the information resource or environment of which it is a part. The methods used derive largely from **ethnography**. The investigators immerse themselves physically in the environment where the information resource is or will be operational, and they collect data primarily through observations, interviews, and reviews of documents. The designs—the data-collection plans—of these studies are not rigidly predetermined and do not unfold in a fixed sequence. They develop dynamically and nonlinearly as the investigators' experience accumulates.

Although subjectivist approaches may run counter to common ideas of how we ought to conduct empirical investigations, these methods and their conceptual underpinnings are not altogether foreign to the worlds of information and computer science. The pluralistic, nonlinear thinking that underlies subjectivist investigation shares many features with modern conceptualizations of the information-resource design process. For example, Winograd and Flores (1987, p. 170) argued as follows:

> In designing computer-based devices, we are not in the position of creating a formal "system" that covers the functioning of the organization and the people within it. When this is attempted, the resulting system (and the space of potential action for people within it) is inflexible and unable to cope with new breakdowns or potentials. Instead we design additions and changes to the network of equipment (some of it computer based) within which people

> work. The computer is like a tool, in that it is brought up for use by people engaged in some domain of action. The use of the tool shapes the potential for what those actions are and how they are conducted. Its power does not lie in having a single purpose … but in its connection to the larger network of communication (electronic, telephone, paper-based) in which organizations operate.

Another connection is to the methodology of **formal systems analysis**, which is generally accepted as an essential component of information resource development. Systems analysis uses many methods that resemble closely the subjectivist methods for evaluation that we introduce here. People recognize that systems analysis requires a process of information gathering, heavily reliant on interviews with people who use the existing system in various ways. Information gathering for systems analysis is typically portrayed as a cyclic, iterative process rather than as a linear process (Davis, 1994). In the literature of systems analysis, we find admonitions, analogous to those made by proponents of subjectivist evaluation, that an overly structured approach can misportray the capabilities of workers in the system's environment, misportray the role of informal communication in the work accomplished, underestimate the prevalence of exceptions, and fail to account for political forces within every organization that shape much of what actually happens (Bansler and Bødker, 1993). Within the field of systems analysis, then, there has developed an appreciation of some of the shortcomings of objectivist methods and of the potential value of subjectivist methods (Zachary et al., 1984).

11.6.2 A Rigorous, but Different, Methodology

The subjectivist approaches to evaluation, like their objectivist counterparts, are empirical methods. Although it is easy to focus only on their differences, these two broad classes of evaluation approaches share many features. In all empirical studies, for example, evidence is collected with great care; the investigators are always aware of what they are doing and why. The evidence is then compiled, interpreted, and ultimately reported. Investigators keep records of their procedures, and these records are open to audit by the investigators themselves or by individuals outside the study team. The principal investigator or evaluation-team leader is under an almost sacred scientific obligation to report their methods. Failure to do so will invalidate a study. Both classes of approaches also share a dependence on theories that guide investigators to explanations of the observed phenomena, as well as to a dependence on the pertinent empirical literature such as published studies that address similar phenomena or similar settings. In both approaches, there are rules of good practice that are generally accepted; it is therefore possible to distinguish a good study from a bad one.

There are, however, fundamental differences between objectivist and subjectivist approaches. First, subjectivist studies are **emergent** in design. Objectivist studies typically begin with a set of hypotheses or specific questions, and with a plan for addressing each member of this set. The investigator assumes that, barring major unforeseen developments, the plan will be followed exactly. Deviation, in fact, might introduce bias. The investigator who sees negative results emerging from the exploration of a particular question or use of a particular measurement instrument might change strategies in hope of obtaining more positive findings. In contrast, subjectivist studies typically begin

with general **orienting issues** that stimulate the early stages of investigation. Through these initial investigations, the important questions for further study emerge. The subjectivist investigator is willing, at virtually any point, to adjust future aspects of the study in light of the most recent information obtained. Subjectivist investigators tend to be **incrementalists**; they change their plans from day-to-day and have a high tolerance for ambiguity and uncertainty. In this respect, they are much like good software developers. Also like software developers, subjectivist investigators must develop the ability to recognize when a project is finished, when further benefit can be obtained only at too great a cost in time, money, or work.

A second feature of subjectivist studies is a **naturalistic** orientation, a reluctance to manipulate the setting of the study, which in most cases is the environment in to which the information resource is introduced. They do not alter the environment to study it. Control groups, placebos, purposeful altering of information resources to create contrasting interventions, and other techniques that are central to the construction of objectivist studies typically are not used. Subjectivist studies will, however, employ quantitative data for descriptive purposes and may offer quantitative comparisons when the research setting offers a "natural experiment" where such comparisons can be made without deliberate intervention. For example, when physicians and nurses both use a clinical system to enter orders, their experiences with the system offer a natural basis for comparison. Subjectivist researchers are opportunists where pertinent information is concerned; they will use what they see as the best information available to illuminate a question under investigation.

A third important distinguishing feature of subjectivist studies is that their end product is a report written in narrative prose. These reports may be lengthy and may require significant time investment from the reader; no technical understanding of quantitative research methodology or statistics is required to comprehend them. Results of subjectivist studies are therefore accessible—and may even be entertaining—to a broad community in a way that results of objectivist studies are not. Objectivist study reports often can be results of inferential statistical analyses that most readers will not find easy to read and will typically not understand. Reports of subjectivist studies seek to engage their audience.

11.6.3 Natural History of a Subjectivist Study

As a first step in describing the methodology of subjectivist evaluation, Figure 11.5 illustrates the stages or natural history of a study. These stages constitute a general sequence, but, as we mentioned, the subjectivist investigator must always be prepared to revise his thinking and possibly return to earlier stages in light of new evidence. Backtracking is a legitimate step in this model.

1. *Negotiation of the ground rules of the study*: In any empirical research, and particularly in evaluation studies, it is important to negotiate an understanding between the study team and the people commissioning the study. This understanding should embrace the general aims of the study; the kinds of methods to be used; the access to various sources of information, including health care providers, patients, and various

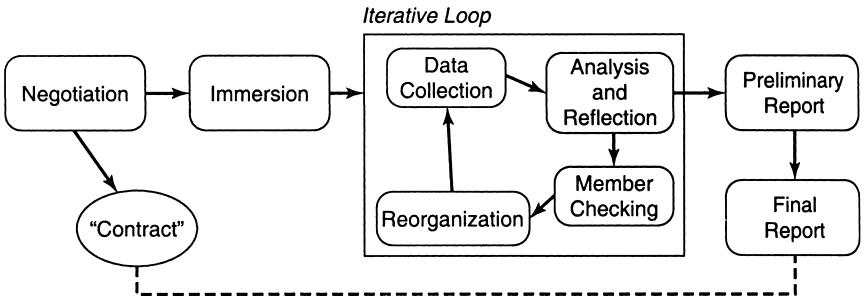

FIGURE 11.5. Natural history of subjectivist studies. (*Source*: Friedman and Wyatt, 1997a.)

documents; and the format for interim and final reports. The aims of the study may be formulated in a set of initial **orienting questions**. Ideally, this understanding will be expressed in a **memorandum of understanding**, analogous to a contract.

2. *Immersion into the environment*: At this stage, the investigators begin spending time in the work environment. Their activities range from formal introductions to informal conversations, or to silent presence at meetings and other events. Investigators use the generic term **field** to refer to the setting, which may be multiple physical locations, where the work under study is carried out. Trust and openness between the investigators and the people in the field are essential elements of subjectivist studies to ensure full and candid exchange of information.

   Even as immersion is taking place, the investigator is already collecting data to sharpen the initial questions or issues guiding the study. Early discussions with people in the field, and other activities primarily targeted toward immersion, inevitably begin to shape the investigators' views. Almost from the outset, the investigator is typically addressing several aspects of the study simultaneously.

3. *Iterative loop*: At this point, the procedural structure of the study becomes akin to an iterative loop, as the investigator engages in cycles of data collection, analysis and reflection, member checking, and reorganization. Data collection involves interview, observation, document analysis, and other methods. Data are collected on planned occasions, as well as serendipitously or spontaneously. The data are recorded carefully and are interpreted in the context of what is already known. Analysis and reflection entail the contemplation of the new findings during each cycle of the loop. **Member checking** is the sharing of the investigator's emerging thoughts and beliefs with the participants themselves. Reorganization results in a revised agenda for data collection in the next cycle of the loop.

   Although each cycle within the iterative loop is depicted as linear, this representation is misleading. Net progress through the loop is clockwise, as shown in Figure 11.5, but backward steps are natural and inevitable. They are not reflective of mistakes or errors. An investigator may, after conducting a series of interviews and studying what participants have said, decide to speak again with one or two participants to clarify their positions on a particular issue.

4. *Preliminary report*: The first draft of the final report should itself be viewed as a research instrument. By sharing this report with a variety of individuals, the investigator obtains a major check on the validity of the findings. Typically, reactions to the preliminary report will generate useful clarifications and a general sharpening of the study findings. Because the report usually is a narrative, it is vitally important that it be well written in language understandable by all intended audiences. Circulation of the report in draft can ensure that the final document communicates as intended. Use of anonymous quotations from interviews and documents makes a report highly vivid and meaningful to readers.
5. *Final report*: The final report, once completed, should be distributed as negotiated in the original memorandum of understanding. Distribution is often accompanied by "meet the investigator" sessions that allow interested persons to ask the author of the report to expand or explain what has been written.

11.6.4 Data-Collection and Data-Analysis Methods

What data-collection strategies are in the subjectivist researcher's black bag? There are several, and they are typically used in combination. We shall discuss each one, assuming a typical setting for a subjectivist study in medical informatics, the introduction of an information resource into patient care activities in a hospital.

Observation

The investigators typically immerse themselves into the setting under study in one of two ways. The investigator may act purely as a detached observer, becoming a trusted and unobtrusive feature of the environment but not a participant in the day-to-day work and thus reliant on multiple "informants" as sources of information. True to the naturalistic feature of this kind of study, great care is taken to diminish the possibility that the presence of the observer will skew the work activities that occur or that the observer will be rejected outright by the ward team. An alternative approach is participant observation, where the investigator becomes a member of the work team. Participant observation is more difficult to engineer; it may require the investigator to have specialized training in the study domain. It is time consuming but can give the investigator a more vivid impression of life in the work environment. During both kinds of observation, data accrue continuously. These data are qualitative and may be of several varieties: statements by health care providers and patients, gestures and other nonverbal expressions of these same individuals, and characteristics of the physical setting that seem to affect the delivery of health care.

Interviews

Subjectivist studies rely heavily on interviews. Formal interviews are occasions where both the investigator and interviewee are aware that the answers to questions are being recorded (on paper or tape) for direct contribution to the evaluation study. Formal interviews vary in their degree of structure. At one extreme is the **unstructured interview**,

where there are no predetermined questions. Between the extremes is the **semistructured interview**, where the investigator specifies in advance a set of topics that he would like to address but is flexible as to the order in which these topics are addressed, and is open to discussion of topics not on the prespecified list. At the other extreme is the **structured interview**, with a schedule of questions that are always presented in the same words and in the same order. In general, the unstructured and semistructured interviews are preferred in subjectivist research. Informal interviews—spontaneous discussions between the investigators and members of a ward team that occur during routine observation—are also part of the data collection process. Informal interviews are invariably considered a source of important data.

Document and Artifact Analysis

Every project produces a trail of papers and other artifacts. These include patient charts, the various versions of a computer program and its documentation, memoranda prepared by the project team, perhaps a cartoon hung on the office door by a ward clerk. Unlike the day-to-day events of patient care, these artifacts do not change once created or introduced. They can be examined retrospectively and referred to repeatedly, as necessary, over the course of a study. Also included under this heading are **unobtrusive measures**, which are the records accrued as part of the routine use of the information resource. They include, for example, user trace files of an information resource. Data from these measures are often quantifiable.

Anything Else that Seems Useful

Subjectivist investigators are supreme opportunists. As questions of importance to a study emerge, the investigators will collect any information that they perceive as bearing on these questions. This data collection could include clinical chart reviews, questionnaires, tests, simulated patients, and other methods more commonly associated with the objectivist approaches.

There are many procedures for analysis of qualitative data. The important point is that the analysis is conducted systematically. In general terms, the investigator looks for themes or trends emerging from several different sources. He collates individual statements and observations by theme, as well as by source. Some investigators transfer these observations to file cards so they can be sorted and resorted in a variety of ways. Others use software especially designed to facilitate analysis of qualitative data (Fielding and Lee, 1991). Because they allow electronic recording of the data will the investigator is "in the field", palm-tops and other hand-held devices are changing the way subjectivist research is carried out.

The subjectivist analysis process is fluid, with analytic goals shifting as the study matures. At an early stage, the goal is primarily to focus the questions that themselves will be the targets of further data elicitation. At the later stages of study, the primary goal is to collate data that address these questions. Conclusions derive credibility from a process of "triangulation", which is the degree to which information from different independent sources generate the same theme or point to the same conclusion.

Subjectivist analysis also employs a strategy known as "member checking" whereby investigators take preliminary conclusions back to the persons in the setting under study, asking if these conclusions make sense, and if not, why not. In subjectivist investigation, unlike objectivist studies, the agenda is never completely closed. The investigator is constantly on the alert for new information that can require a significant reorganization of the findings and conclusions that have been drawn to date.

1.7 Conclusions: The Mindset of Evaluation and Technology Assessment

The previous sections probably make evaluation and technology assessment look difficult. If scholars of the field disagree in fundamental ways about how these studies should be done, how can relative novices proceed at all, much less with confidence? To address this dilemma, we conclude this chapter by offering a mindset for evaluation, a general orientation that anyone conducting an evaluation might constructively bring to their work. The components of this mindset apply, to varying degrees, across all study types and approaches.

- *Tailor the study to the problem and key stakeholder questions*: Every study is made to order. Evaluation and technology assessment differ from mainstream views of research in that a study derives importance from the needs of clients rather than from the unanswered questions of an academic discipline. If an evaluation contributes new knowledge of general importance to an academic discipline, that is a serendipitous by-product.
- *Collect data that will be useful to make decisions*: There is no theoretical limit to the questions that can be asked and, consequently, to the data that can be collected in a study. What is done is determined by the decisions that need ultimately to be made and the information seen as useful to inform these decisions.
- *Look for intended and unintended effects*: Whenever a new information resource is introduced into an environment, there can be many consequences, only some of which relate to the stated purpose of the resource. In a complete evaluation, it is important to look for and document effects that were anticipated as well as those that were not, and to continue the study long enough to allow these effects to manifest themselves.
- *Study the resource while it is under development and after it is installed*: In general, the kinds of decisions evaluation can facilitate are of two types. **Formative decisions** are made as a result of studies undertaken while a resource is under development. They affect the resource before it can go online. **Summative decisions** are made after a resource is installed in its envisioned environment and deal explicitly with how effectively the resource performs in that environment. Often, it will take many years for an installed resource to stabilize within an environment. Before summative studies are conducted, it may be necessary for this amount of time to pass.
- *Study the resource in the laboratory and in the field*: Completely different questions arise when an information resource is still in the laboratory and when it is in the field. *In vitro* studies, conducted in the developer's laboratory, and *in vivo* studies, conducted

in an ongoing clinical or educational environment, are both important aspects of evaluation.

- *Go beyond the developer's point of view*: The developers of an information resource usually are empathic only up to a point and are often not predisposed to be detached and objective about the resource's performance and utility. People doing evaluation often see it as part of their job to get close to the end user and to portray the resource as the user sees it.
- *Take the environment into account*: Anyone who conducts an evaluation study must be, in part, an ecologist. The function of an information resource must be viewed as an interaction among the resource itself, a set of users of the resource, and the social, organizational, and cultural context that largely determines how work is carried out in that environment. Whether a new resource functions effectively is determined as much by its goodness of fit with its environment as by its compliance with the resource designers' operational specifications as measured in the laboratory.
- *Let the key issues emerge over time*: Evaluation studies are dynamic. The design for a study, as it might be stated in a project proposal, is often just a starting point. Rarely are the important questions known, with total precision or confidence, at the outset of a study. In the real world, evaluation designs, even those employing objectivist approaches, must have some leeway to evolve as the important issues come into focus.
- *Be methodologically Catholic and eclectic*: It is best to derive overall approaches, study designs, and data collection methods from the questions to be explored rather than to bring predetermined methods or instruments to a study. Certain questions are better answered with qualitative data collected through open-ended interviews and observation. Others are better answered with quantitative data collected via structured questionnaires, patient chart audits, and logs of user behavior.

Finally, remember that the perfect study has never been performed and probably never will be. This chapter has introduced various approaches to study design and execution that can minimize bias and maximize credibility, but the findings of every study can be questioned. It is sufficient for a study to be guiding, clarifying, or illuminating.

Suggested Readings

Anderson J.G., Aydin C.E., Jay S.J. (Eds.) (1994). *Evaluating Health Care Information Systems: Methods and Applications*. Thousand Oaks, CA: Sage Publications.

This is an excellent edited volume that covers a wide range of methodological and substantive issues in evaluation, including both objectivist and subjectivist approaches. Although not formally constructed as a textbook, it is written at a basic level for individuals more familiar with medical informatics than study methodology.

Cohen P.R. (1995). *Empirical Methods for Artificial Intelligence*. Cambridge, MA: MIT Press.

This is a nicely written, detailed book that is focused on evaluation of artificial intelligence applications, not necessarily those operating in medical domains. It emphasizes objectivist methods and could serve as a basic statistics course for computer science students.

Friedman C.P., Wyatt J.C. (1997). *Evaluation Methods in Medical Informatics*. New York: Springer-Verlag.
This is the book on which the current chapter is based. It offers expanded discussion of almost all issues and concepts raised in the current chapter.

Jain R. (1991). *The Art of Computer Systems Performance Analysis: Techniques for Experimental Design, Measurement, Simulation, and Modeling*. New York: John Wiley & Sons.
This work offers a technical discussion of a range of objectivist methods used to study computer systems. The scope is broader than Cohen's book (1995) described earlier. It contains many case studies and examples and assumes knowledge of basic statistics.

Lincoln Y.S., Guba E.G. (1985). *Naturalistic Inquiry*. Beverly Hills, CA: Sage Publications.
This is a classic book on subjectivist methods. The work is very rigorous but also very easy to read. Because it does not focus on medical domains or information systems, readers must make their own extrapolations.

Rossi P.H., Freeman H.E. (1989). *Evaluation: A Systematic Approach* (4th ed.). Newbury Park, CA: Sage Publications.
This is a valuable textbook on evaluation, emphasizing objectivist methods, and is very well written. Like the book of Lincoln and Guba (1985), described earlier, it is generic in scope, and the reader must relate the content to medical informatics. There are several excellent chapters addressing pragmatic issues of evaluation. These nicely complement the chapters on statistics and formal study designs.

Questions for Discussion

1. Choose any alternative area of biomedicine (e.g., drug trials) as a point of comparison, and list at least four factors that make studies in medical informatics more difficult to conduct successfully than in that area. Given these difficulties, discuss whether it is worthwhile to conduct empirical studies in medical informatics or whether we should use intuition or the marketplace as the primary indicators of the value of an information resource.
2. Assume that you run a philanthropic organization that supports medical informatics. In investing the scarce resources of your organization, you have to choose between funding a new system or resource development, or funding empirical studies of resources already developed. What would you choose? How would you justify your decision?
3. To what extent is it possible to be certain how effective a medical informatics resource really is? What are the most important criteria of effectiveness?
4. Do you believe that independent, unbiased observers of the same behavior or outcome should agree on the quality of that outcome?
5. Many of the evaluation approaches assert that a single unbiased observer is a legitimate source of information in an evaluation, even if that observer's data or judgments are unsubstantiated by other people. Give examples drawn from our society where we vest important decisions in a single experienced and presumed impartial individual.

6. Do you agree with the statement that all evaluations appear equivocal when subjected to serious scrutiny? Explain your answer.
7. Associate each of the following hypothetical studies with a particular approach to evaluation:[3]
   a. A comparison of different user interfaces for a computer-based medical-record system, conducted while the system is under development.
   b. A site visit by the U.S. National Library of Medicine's Biomedical Library Review Committee to the submitters of a competing renewal of a research grant.
   c. A noted consultant on user interface design being invited to spend a day at an academic department to offer suggestions regarding the prototype of a new system.
   d. Patient chart reviews conducted before and after the introduction of an information resource, without the reviewer being told anything about the nature of the information resource or even that the intervention is the information resource.
   e. Videotapes of attending rounds on a service where a knowledge resource has been implemented and periodic interviews with members of the ward team.
   f. Determination of whether a new version of a resource executes a standard set of performance tests at the speed the designers projected.
   g. Patients being randomly assigned such that their medical records are maintained either by a new computer system or by standard procedures, and then an investigator seeking to determine whether the new system affects clinical protocol recruitment and compliance.
   h. A mock debate at a research-group retreat.
8. For each of the following hypothetical evaluation scenarios, list which of the eight types of studies in Table 11.2 they include. Some scenarios may include more than one type of study.[4]
   a. An order-communication system is implemented in a small hospital. Changes in laboratory workload are assessed.
   b. A study team performs a thorough analysis of the information required by psychiatrists to whom patients are referred by community social workers.
   c. A medical-informatics expert is asked for opinion about a doctoral student's project. The expert requests copies of the student's programming code and documentation for review.
   d. A new intensive care unit system is implemented alongside manual paper charting for one month. Then, the qualities of the computer-based data and of the data recorded on the paper charts are compared. A panel of intensive care physicians is asked to identify episodes of hypotension from each dataset, independently.
   e. A medical-informatics professor is invited to join the steering group for a clinical-workstation project in a local hospital. The only documentation available for the professor to critique at the first meeting is a statement of the project goals, a

[3]Answers: (a) decision facilitation; (b) professional review; (c) art criticism; (d) goal free; (e) responsive–-illuminative; (f) objectives based; (g) comparison based; (h) quasi-legal.

[4]Answers: (a) clinical impact; (b) need validation; (c) structure validation; (d) field function; (e) design validation and need validation; (f) laboratory function; (g) laboratory user impact, laboratory function; (h) clinical impact.

description of the planned development method, and the advertisements and job descriptions for team members.

f. Developers invite clinicians to test a prototype of a computer-aided learning system as part of a workshop on user-centered design.
g. A program is built that generates a predicted 24-hour blood glucose profile using seven clinical parameters. Another program uses this profile and other patient data to advise on insulin dosages. Diabetologists are asked to prescribe insulin for the patient given the 24-hour profile alone and then again after seeing the computer-generated advice. They are also asked their opinion of the advice.
h. A program to generate drug-interaction alerts is installed in a geriatric clinic that already has a computer-based medical record system. Rates of clinically significant drug interactions are compared before and after installation of the alerting resource.

Author Queries:

[JW1] Chuck, I reckon that Doug contributed about 3.5 pages and I contributed about 7 to the chapter -could we switch 2nd / 3rd author order do you think this time ?
[Q2] Is the deletion OK?
[JW3] not sure this is entriely compatible with the falsification view above !
[JW4] I prefer to shift the focus away form boring systems to people !